Project Report – FY2004

Core Name: Applied Marine Genomics

Project Title: Bioinformatics

Reporting Period: 1 September 2005 – 30 September 2006

Principal Investigators: Robert W. Chapman SCDNR

Associate Investigators: Louis Burnett and Karen Burnett, College of Charleston; Paul Gross Medical University of South Carolina; Jonas Almeida, MD Anderson Cancer Center; Matt Jenny, Woods Hole Oceanographic Institute.

Background and Rationale:

The role of the Bioinformatics Project is to provide advanced computational analyses for data generated by the Applied Genomics Program and to collaborate with the HML Data Management Section to integrate marine genomics data into the overall synthesis and integration activities of the NOAA Center of Excellence for Oceans and Human Health at the HML and the HML in general.

Over the past decade, the explosion in molecular biology and genomic technology has provided biologists the capabilities to produce unparalleled quantities of data about the thousands of genes in individual organisms. Historically, assessment of transcriptional changes in a few genes associated with an environmental challenge was a major undertaking. With this new knowledge and technology, surveying the changes in much of the transcriptome is currently possible for key marine species.

For most genomic studies, the number of independent variables (individual gene expression levels) is huge and the sample sizes (number of individual animals and sites studied) are small. Analysis of these data using conventional statistical approaches produces more unknowns than equations and lack sufficient degrees of freedom to solve the linear matrices. In addition, few phenomena in biology are linear. Therefore, the most appropriate mathematical structure for representing the data is generally unclear. A final complication to the understanding of genetic profiles is that in normal physiological responses, nearly all genes produce products that are part of a metabolic cascade and their individual contributions to metabolic output might be small. However, their collective (epistatic) interactions are likely to be major factors affecting the process and output. While such effects can be identified by linear statistics, they cannot be effectively modeled using this approach.

Analysis and synthesis of the complex data resulting from genomics research, therefore, require new statistical approaches, called bioinformatics. Bioinformatic approaches are drastically different from traditional statistical approaches to data analysis and are based upon the application of non-linear modeling tools (e.g., artificial neural networks) and machine learning to mine the data and information that exist in complex genomic data

sets. This SOW describes development of a small bioinformatics analysis group as part of the Applied Genomics Program for the NOAA Center of Excellence for Oceans and Human Health at the HML. This group will adapt and apply bioinformatics data mining and modeling tools to synthesize and integrate the genomics data resulting for the Applied Genomics Programs and work with the Monitoring and Assessment Core to incorporate genomics data into analyses that assess the ecosystem and public health.

Approach:

The ultimate goal of the bioinformatics program is to generate mathematical representations of gene expression profiles, which can be used to identify new and innovative biomarkers of environmental stress and accomplish the transition from the dynamics of transcript profiles in individuals to the behaviors of populations and ecosystems.

Analysis Approach 1: Pre-filtering via Artificial Neural Network

This approach includes the following steps: (1) training ANN'S to map transcript profiles to the response variable(s) via maximization the correlation between observation and prediction using no more than 90% of the available records and encompassing the full dynamic range of output variables; (2) identifying the sensitivities of response variables to the individual gene expression levels; (3) eliminating the less important or non-contributing transcripts; (4) repeat step one with the remaining inputs; and (5) testing the revised model with the remaining records. We expect, based upon previous experience, that ANN models should produce at least a 90% correlation between observed and expected values of the output variables (contaminant levels, immune status, or *Perkensis* infection status, described in the Functional Genomics SOW)

Analysis Approach 2: Clustering Methods:

This approach uses a modified Pearson correlation coefficient or other metrics to perform two-dimensional clustering that reduces the vast quantity of array data to easily visualized gene trees and treatment trees. A number of variations on this general theme are available including a range of metrics on which to apply classification or clustering algorithms (e.g., http://rana.lbl.gov/EisenSoftware.htm).

We will employ clustering to identify genes that are up and down regulated by particular stressors and/or groups of stressor for input into the ANN analysis. Again, some portion of the data set will be withheld for subsequent ANN testing and validation.

Analysis Approach 3: Statistical Analysis Using Rank Correlations:

Recently, the Marine Genomics Group developed an alternative analysis approach to identify up- and down-regulated genetic responses in microarrays (prefiltering) that employs a Spearman rank correlation coefficient. This approach begins by rank ordering the level of expression for each gene on the microarray and normalizing the ranks between 0 and 1. The most highly expressed sequence in the sample is given the rank 1.0, and genes which are not expressed are given the rank 0. The scores of remaining sequences are given a normalized score between 0 and 1 based on their rank order. The next step is calculation of rank order correlations among samples and comparison of rank order correlations between control individuals including the estimation of confidence intervals (usually 95%). The confidence intervals are then overlaid upon correlation maps

of control versus experimental animals. Genes lying outside the confidence interval are selected for further analysis. Those above the 95% confidence interval are assumed to be up-regulated and those below the confidence limits are assumed to be down regulated. This method has several advantages over other approaches including: (1) no reliance upon arbitrary values to determine up- or down- regulation and (2) variations in signal strength among arrays (due to labeling mRNA, hybridization conditions, etc.) is effectively neutralized.

The singular advantage to using alternative methods for gene identification (Cluster and Spearman) over ANN sensitivities is a pure matter of speed. Our experience suggests the ANN pre-filtering process will require 1-4 days of computation time for each output variable on existing 32 bit processors. Most classification approaches and the correlation method can be accomplished in a matter of milliseconds. The disadvantage is that low-level changes in transcription profiles and genetic signatures, which may have important consequences for physiological responses and adaptation, are likely to be missed. Nonetheless, it is important that the comparisons be made to determine how much information is lost using these linear short cuts.

Regardless of the data set and analysis approach, it will be important to identify which of the vast number of genes on the microarray are truly responding to environmental challenges of interest. This is necessary for several reasons. First, the complex microarray can provide information on the genetic response (transcriptome) for 1000s of genes. They are, however, expensive and time-consuming to develop and may provide more information than can be interpreted or comprehended given the present of state-of-knowledge for most marine species. If a microarray can be reduced to a relatively small number of genes that represents the overall response of the individual to the challenge (e.g., 100s of genes), then the biological significance of the data will be easier to comprehend and apply. Reducing the number of genes being evaluated would also shorten the computation time and reduce the hardware resources needed. There are a variety of approaches to accomplish the task of transcribing a microarray comprised of 1000s of genes to a much smaller suite that best represents the overall organism response: clustering and self-organizing maps (Eisen *et al.* 1998), shrinkage-based metrics (Cherepinsky, et al. 2003), and Spearman rank correlations, as well as sensitivity analyses from the ANN itself.

Objectives:
The objectives of the Bioinformatics Project are to:

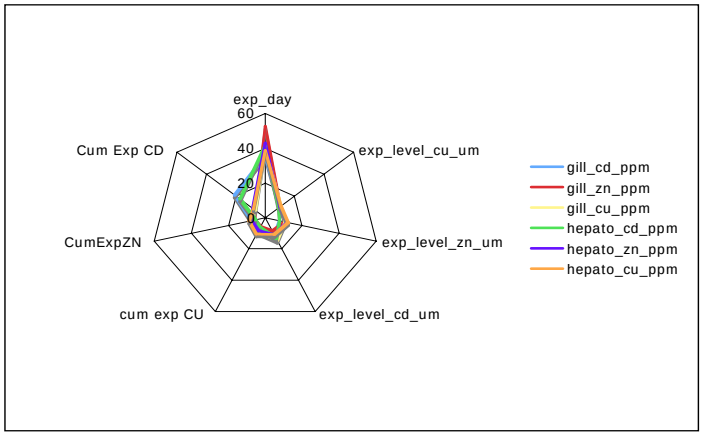
1. Apply non-linear approaches to develop mathematical models of gene expression profiles generated by the Functional Genomics Research and Marine Organisms as Disease Vectors projects for the oysters, *Crassostrea virginica* and *C. gigas,* and their pathogen, *Perkinsus marinus.*

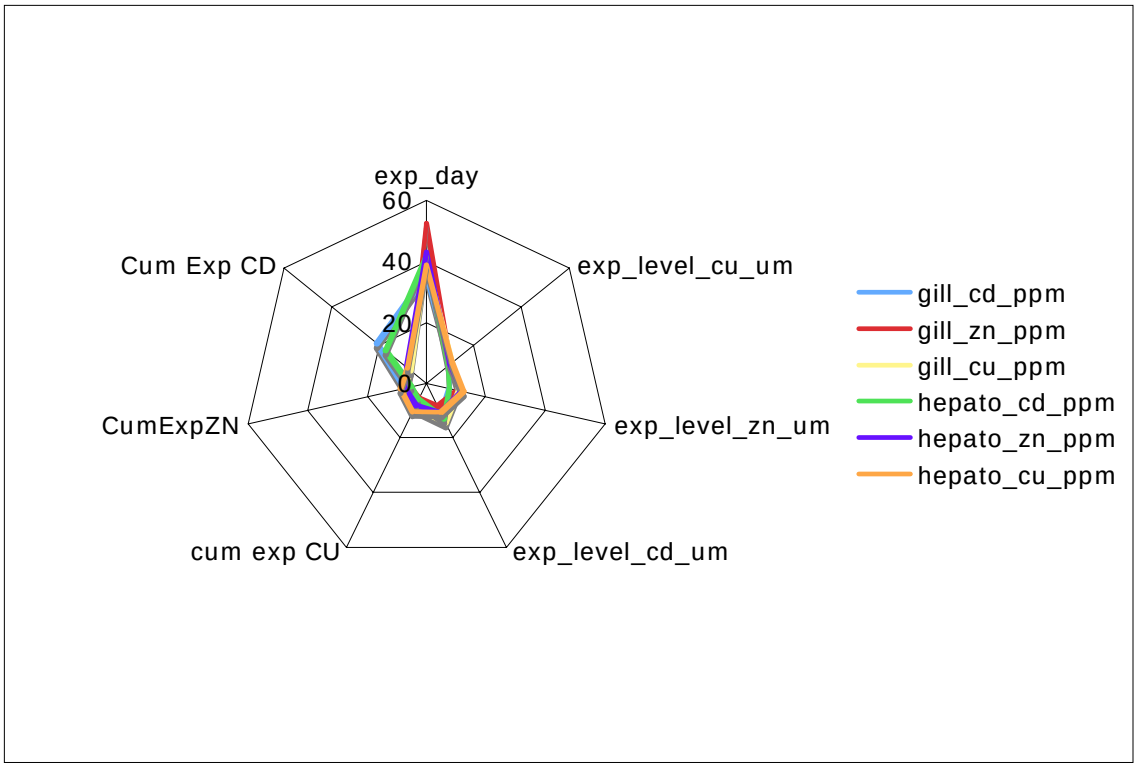
2. Develop similar analysis models for mapping transcript profiles of oysters to the other environmental data collected by the Monitoring and Assessment Core of the NOAA Center of Excellence for Oceans and Human Health at the HML.

3. Generate lists of genes from 1 and 2 above that appear to be significantly up or down regulated by the environmental conditions to which the oysters have been exposed

Accomplishments:

The BioInoformatic core has just received the necessary microarray data to begin working with the large data set generated by the Functional genomics section. These data have been examined for relationships between environmental exposure levels on the accumulation in tissues of three metals Cd. Cu and Zn. The results indicate that there is little correlation between exposure and accumulation of Cu and Zn, but a strong correlation exits for Cd. However, the Cd accumulation in tissues is dependent more on the length of exposure than on environmental concentrations. In the figure to the right, the sensitivities of the concentration of the three metals in gill and hepatopancreas tissues to exposure time and exposure levels are illustrated. These data were generated by 30 runs of an Artificial Neural Network and exposure time accounted for more than 40% of the variation in tissue concentrations. Further analysis of these were conducted by withholding 20 samples from each ANN run and then using the models to predict the tissue concentration of each metal. In the figure below the average correlations between the observed data and predictions are shown. Here, it is evident that only Cd can be predicted with reasonable accuracy.





In a related experiment conducted by Drs. Matt Jenny and Charlie Cunningham, oysters were challenged by a bacterial treatment and tissue samples taken from microarray analysis. These microarrays were compared to those from notched individuals and saline injected individuals. In the analysis we developed an approach for reducing the number of transcripts necessary to discriminate individuals in each treatment. The process began by assessing the sensitivity of discrimination for each transcript with a single pass through an ANN. The highest ranking transcripts (200) were taken for an addition 30 runs of the program for each treatment class. In the figure below a Receiver Operating Characteristic curve (ROC) demonstrates that correct classification of individuals subjected to bacterial challenge (24 hr post injection) can be achieved 94% of the time. False positives and false negatives comprise less than 6% of the data.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

In addition to providing accurate predictions of bacterial challenge, this procedure reduced the number of transcripts to models the outcomes from 7000 to 200 and accelerated the processing time from days to hours with little or no degradation in the signal.

With the assistance of Dr. Jonas Almeida, the Marine Genomics Staff has developed a web accessible Microarray pipeline. The tool filters "bad" spots on the replicate Microarrays, normalizes the quantiles and spots, compares the genes on each Microarray to determine differential gene expression and returns a cumulative list of genes categorized by their level of regulation in the presence of stressos. Registered curators can analyze their data using the following statistical functions made available to them:

- Array data easy retrieval to MATLAB
- Array 'bad-spot' filtering
- Within-array calibration
- Spearman/Pearson values
- Differential expression values determination
- Array clustering
- Excel report generation of all values including differential values, p-values, fold difference values between spots across arrays etc.
- graphical reports
- Principle component analysis

Publications:

Robert W. Chapman. Javier Robalino, and Hal Trent. (2006) EcoGenomics: Analysis of Complex Systems Using Fractal Geometry. Integrative and Comparative Biology. Published on line May 2006

Presentations:

Chapman RW, McKillen DJ, Trent HF, Chen YA, Almeida JS, Gross PS, Warr GW, Robalino J, Jenny M, Cunningham C. Ecogenomics: Analytical Challenges and Potential Solutions. Tenth Congress of the International Society for Developmental and Comparative Immunology, Charleston, SC, July 2006

Application/Technology Transfer:

1.0 Scientific Research and Application

The bioinformatics developments at HML are breaking new ground in the analysis of microarray data. In addition the ANN analysis tools can be applied to all types of data and provide new and compelling insights into biological processes. The development of web based access to these models is under development.

2.0 Public Information and Outreach

The marine genomics website provides a publicly accessible database and set of analytical tools for understanding the genomics of selected marine organisms (including the American oyster, Atlantic white shrimp, grass shrimp and the Atlantic bottlenose dolphin) and assessing their health using transcriptomic methods. In addition the models developed from the analysis of the oyster data will become available over our web site in the future and accessible to researchers wishing to use them.

3.0 Capacity Building

Additional computational power has been acquired over the past few months, but even with these resources we are taxing our system. Nonetheless these servers are available over the web for processing DNA sequence data, assembling contig and cluster information and access to microarray data.

Project Abstract:

The basic tools and methods have been developed for interrogating microarray data to extract the important signals and predict organism responses to a variety of stressors. In addition insights into the biology of metal accumulation in oyster and the value of various measure of toxicity have been explored.

Unresolved issues:

- The bioinformatics capacity (in terms of both personnel and computing capacity) to manage and analyze extremely large datasets continues to be a major challenge for the Functional Marine Genomics Projects that must be overcome. A major challenge identified in the previous Progress Report, i.e. the translation of the research from laboratory-scale populations to broad, living ecosystems is currently being addressed in a direct test of oysters sampled from field sites.

Budget Report:

As of 31 August 2006, expenditures and commitments totaled:

	Year 1	Year 2	Year 3
Salaries	22,667.50	22,667.50	43,024
Fringe	7,480.50	7,480.50	22,860
Supplies	24,000.00	24,000.00	0
Contractual	0.00	0.00	0
Indirect	3,305.00	3,305.00	1,105
Total	57,453.00	57,453.00	67,089